Analysis of Customer Profiles on an Electrical Distribution Network

Brian Akperi
School of Engineering and
Computing Sciences
Durham University
Durham, United Kingdom
Email: b.t.akperi@durham.ac.uk

Peter Matthews
School of Engineering and
Computing Sciences
Durham University
Durham, United Kingdom
Email: p.c.matthews@durham.ac.uk

***Abstract*—It has become increasingly important for electrical distribution companies to understand the drivers of demand. The maximum demand at any given substation can vary materially on an annual basis which means it is difficult to create a load related investment plan that is robust and stable. Currently, forecasts are based only on historical demand with little understanding about contributions to load profiles. In particular, the unique diversity of customers on any particular substation can affect load profile shape and future forecasts. Domestic and commercial customers can have very different behaviours generally and within these groups there is room for variation due to economic conditions and building types.**

This paper analyses customer types associated to substations on a distribution network by way of principal component analysis and identification of substations which deviate from the national demand trend. By examining the variance spread of this deviation, data points can be labelled in the principal component space. Groups of substations can then be categorised as having typical or atypical load profiles. This will support the need for further investigation into particular customer types and highlight the key factors of customer categorisation.

I. Introduction

In order to have a better understanding of trends in substation load profiles, it is necessary to investigate the customer make-up at primary substations at the distribution level. Electric load demand is collected using SCADA (Supervisory Control and Data Acquisition) systems at 30 minute intervals. This data is collected for substations typically on 11kV or 33kV distribution and a maximum demand is calculated based on this data. There are several uncertainties associated to this, one of them being the customer usage behaviour. In particular, there is a distinction between domestic and commercial customers and their contribution to substation demand.

Customer data has been collected by the sponsor Northern Powergrid for their primary substations. The initial categorisation is by postal sector which can then be associated to a primary substation in that area. The two major headings are domestic houses and commercial buildings both of which are further split into subcategories. There are 20 domestic house types and 15 commercial building types for a total of 35 subcategories. This high dimensionality makes it difficult to understand the impact of individual contributions. Thus a combined approach of principal component analysis with input from national demand trend using clustering will be used to determine load profile types worth greater investigation.

This demographic information that specifies the type of customer associated to each substation is usually not readily available so analysis of this type of data is not often seen. There are several studies that perform customer classification based mostly on load patterns without a preliminary demographic study [1] [2]. This investigation aims to determine whether an in-house classification of customers is useful at the distribution level.

National demand data is available from the National Grid on a half hourly scale [3]. This represents the prototypical load profile for the country accounting for national trends in weather and economic conditions. Substations in a distribution network on a local scale are not expected to follow this pattern but deviation from this trend can logically be attributed to the individual customer make-up of the substation.

This paper will use a combined approach of principal component analysis and clustering to reduce the dimensionality of the domestic and commercial descriptors and categorise substations based on their loadings. In particular, a principal component biplot will be used to give a visualisation of the customer categorisations and their relationship to each other.

II. Related Work

A detailed customer breakdown on the distribution network is available that would normally would not be available in a general study. The need for this type of information stems from the SCADA data being too noisy to offer engineers detailed insight. There is a financial incentive to understand the data such as tariff determination for energy suppliers but also a more general understanding is needed for reliability analysis. In the literature, data mining techniques such as clustering and classification algorithms are popular as in [2]. Prevalent in both [1] and [2] is the concept of obtaining representative

sets of load profiles based on customer and meteorological conditions. Work done in [4] presents a framework that also uses a data mining approach but with additional indices for classification based on the time of day. These data mining methods are all able to offer customer categorisation methods for SCADA data with no need for additional information. This study will be unique in the synthesizing of in-house customer classifications with a novel trend identification method.

III. Data Set Description

There are three main data sets used in this investigation. First is the customer breakdown data commissioned by NPG which associates to each NPG substation, the number of customers in 35 different categories, 20 of which are domestic houses and 15 of which are commercial building types.

The domestic housing types are given a label 1-20 based on five descriptors: fuel, location, size, tenure and age. The fuel descriptor indicates whether or not the house has a mains gas connection or only uses electricity. The location descriptor shows if the house is either in a rural or urban area. The size descriptor shows if the house is large or small where a small house is generally a flat. The tenure descriptor indicates whether the house is social (council) housing or not. The age descriptor of the house is given by pre-selected time periods. New buildings (sparsely populated) are defined as those built within the last year of the data compilation, a recent house is one built post 1980, old houses from 1920-1979 and houses pre-1920 as very old.

The commercial customer types only have their descriptors available which are *a*) Business at Home *b*) Shops & Other Retail Outlets *c*) Sports, Leisure, Entertainment, Holiday Activities *d*) Unknown *e*) Warehouses and Wholesalers *f*) Office and Administration *g*) Head Office *h*) Other *i*) Schools & Educational Establishments *j*) Transport *k*) Workshops & Repair Centres *l*) Factories & Manufacturing *m*) Hospital & Medical Establishment *n*) Places of Worship and *o*) Police, Fire, Ambulance, Courts, Prisons, Civil Defence, Libraries.

Second, the half hourly demand data for 513 substations is available from a period of April 2010 - March 2013. Some substations cannot be attributed to the customer database because of missing information so this is reduced to 436 substations in total.

Third, the half hourly demand data publicly available from the National Grid from April 2010 - March 2013 which represents the prototypical load profile for the country.

IV. Theory and Methods Used

A. PCA Overview

The aim of principal component analysis is to analyse multivariate data by way of reducing the dimensionality of the data set. This is done by transforming the original space into a new set of uncorrelated variables in decreasing order of importance. The first principal component (PC) retains most of the variance in the original data set, the second PC retains the second most variance and so on.

The principal component space can also be rotated to help with interpretation. After picking a lower dimensional space from PCA, the space can be rotated so that the loadings fall closer to the principal component axes. In general, the principal component space can be rotated either by an orthogonal transformation or an oblique one [5]. Suppose the original matrix of loadings is given by A_m, a $35 \times m$ matrix in this investigation where m is a chosen number of dimensions. For an orthogonal transformation, a rotation matrix T is found such that a new matrix $B_m = A_m T$ of loadings optimises a simplicity criteria [5].

Analysing relationships between variables in the original space can be difficult just by examining a covariance matrix and similarly the principal component scores of the transformed data. Therefore it is useful to have a graphical representation of the data being examined. The purpose of the biplot is threefold. One is to plot the PC scores in a 2D (or 3D) space so that relationships between individual substations can be deduced in the new space. Second is to determine where the original variables lie in the new PC space and their relation to each other. Finally, the simultaneous position of both the data point PC scores and original variables represented by vectors can be considered. The direction of the vectors can give an indication of a point having a high or low PC score in that variable.

B. K-Means Clustering Overview

K-means clustering is known as an unsupervised learning algorithm because its objective is not to predict a classification but rather to find patterns in data without labels. After specifying the number of desired clusters, points are chosen at random to be cluster centres. All points are assigned to their closest centre by a distance metric, typically Euclidean distance. Then the means of all points in each cluster are taken and these become the new centres. The algorithm stops when all points are assigned to the same cluster in consecutive rounds [6].

The initialisation of points chosen as cluster centres can prove to be problematic as they can often result in different clusters. In order to obtain a good solution, the clustering algorithm will be ran 100 times and the solution with the lowest total squared distance between the substation points and the centroids will be chosen [6].

There is evidence to support that when using k-means clustering with the Euclidean distance metric that using the first few principal components yields accurate clusters [7]. In order to determine the number of clusters, the Davies-Bouldin (DB) criterion will be used which is a ratio of within

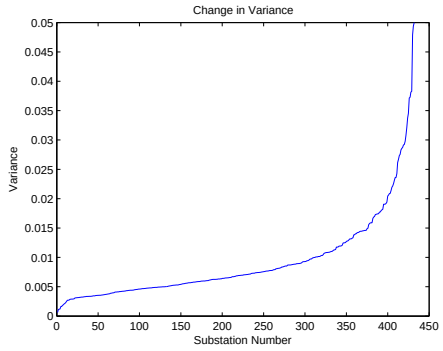


Fig. 1. Change in Error Variance between Substation Trend and National Trend

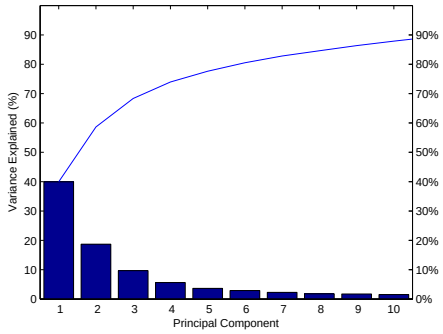


Fig. 2. Scree Plot of Variance Explained by Principal Components

cluster and between cluster differences. Suppose S_i and S_j are dispersion measures which are the average distances between each point in the clusters and their respective centroids. $M_{i,j}$ is the Euclidean distance between the ith and jth clusters. Then the DB index $\bar{R}$ is

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} max_{i\neq j}\{R_{i,j}\},\ R_{i,j} = \frac{S_i + S_j}{M_{i,j}}. \qquad (1)$$

The best clustering solution has the smallest DB index [8].

C. Comparing Monthly Growths

In order to compare the monthly growths of the demand with the National Grid demand, all of the substation demands and national demand from April 2010 - March 2013 are normalised. A comparison can be achieved by finding the difference between the monthly changes and then calculating the variance to show the spread of errors. The change in variance from the lowest error substation to the highest is shown in Fig. 1.

A decision regarding the choice of variance value where load profiles are said to more closely follow the national trend is nontrivial. The point to be selected as a threshold value needs to select a subset of the data that not only includes outliers but also includes those points which start to become more sparsely spread out. In Section VI, several threshold values are looked at but for now, 0.01 is selected as a threshold value because variance points become more spread out and it leaves out about a quarter of the data as being the furthest away from the national trend.

V. Usage of Principal Component Analysis

Principal component analysis was applied to the customer profiles of the 436 substations in the distribution network. A scree plot of the variance explained by the first 10 principal components is given in Fig. 2. A popular method for determining the number of principal components that should be used is looking at where the scree plot starts to level out. For the purposes of visualising the PC space, the first three PCs are selected and this is also justified by the variance explained in the scree plot. The first three PCs explain about 71% of the variance in the original space.

This data was also plotted in Matlab using the biplot function along with the scores for each of the 436 substations. By convention, the biplot function makes the element with the largest absolute value in each column of the factor loadings matrix a positive value by changing the sign on the entire axes. In this case, PC2 was mirrored. This does not change the meaning of the plot. Also, the scores of the actual substations are scaled down so that they will fit the plot.

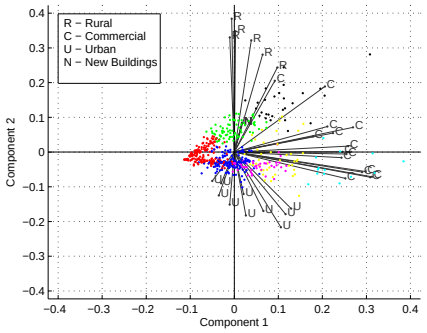


Fig. 3. K-Means Clustering of PC Space

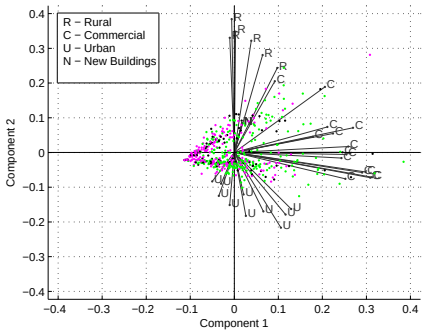
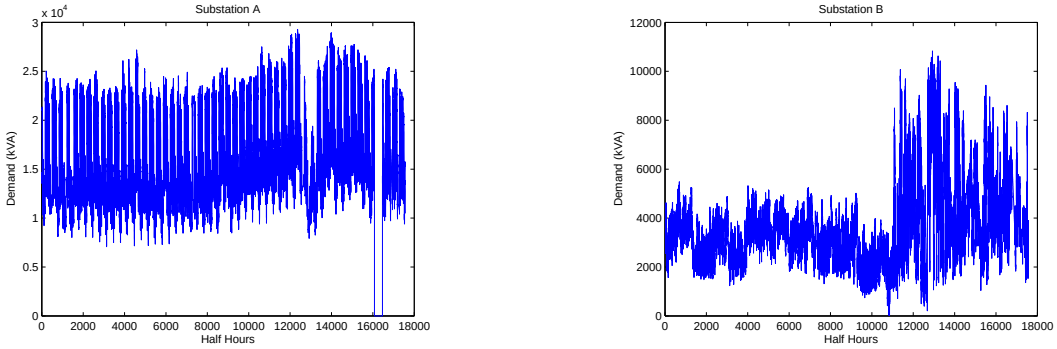


Fig. 4. Substations which Deviate from National Trend

TABLE I
SUBSTATIONS THAT MATCHED TREND BY CLUSTER

	1 - Yellow	2 - Magenta	3 - Cyan	4 - Red	5 - Green	6 - Blue	7 - Black
Matches Trend	31	31	11	70	40	110	19
Does not match trend	5	7	0	59	23	25	5



(a) Case I - High PC1 Score

(b) Case II - Low PC1 Score

Fig. 5. Difference between Load Profiles in Principal Component 1

In Fig. 3, it is possible to see a visual representation of the first two principal components (from the 3D space). From here, it is possible to see the correlation between certain categories of customers. In the figure, the customer labels are R for domestic rural housing, U for domestic urban housing, C for commercial customers and N for new domestic buildings. Almost all of the commercial customer classifications are grouped together as well as certain types of domestic customers where the main divide seems to be whether the customer is urban or rural. The scores of the substations themselves seem to be more concentrated towards the left of the plot which suggests there is a greater influence from domestic customers, in particular those in urban environments.

Afterwards, k-means clustering is applied on the three dimensional PC scores and by using the DB criterion, seven clusters are found to be optimal. Fig. 3 shows the clusters in the PC space.

Now using a variance of 0.01 as a threshold value, the substations which fall below and above this can be separated on the same PC space as in Fig. 4. Substations which deviate greatly are in purple and those which do not are in green. By comparing this to the initial k-means clustering of the space, Table I shows the number of substations in each cluster that fall above or below this threshold value. The clear anomaly in this table is the red cluster which is concentrated in the left of the plot. These substations are the ones which most strongly deviate from the commercial variable vectors on the right of the biplot. Another interesting observation is that the substations in the green and blue cluster containing the next highest proportion of substations that deviate from the national trend are still distinct from the red cluster. This suggests that the spread of the urban variables may also be significant. In particular, domestic groups 1, 3 and 5 in the bottom left are all large urban houses built no earlier than 1920 whereas the other urban customers occupy flats or have no mains gas connection. Although this is not enough evidence to dismiss the need for investigating the latter group, it does suggest the former group of customers is worth further investigation.

As an illustration of the power of PCA, Fig. 5 shows the difference between a load profile with a high PC1 score and a load profile with a low PC1 score. The load profile in Case I not only has a higher load demand but is more orderly and levelled. The load profile in Case II is erratic by comparison so it is unsurprising that it would not follow a national trend as closely.

VI. ADJUSTMENT OF THRESHOLD VALUE AND ERROR METRICS

As the choice of threshold parameter is somewhat subjective, there needs to be an allowance for it to be adjusted by distribution network engineers. Using the cluster analysis, it is also possible to assign error metrics to these groupings. In addition to the threshold, there needs to be allowance for adjustment of the proportion (P) of substations in a cluster that define whether it follows the trend or not. Let the clusters which have a proportion of substations greater or equal to P that do not match the national trend be a prediction of a substation not following the national trend. Then by using a confusion matrix, the engineer can gain an understanding of the accuracy of the cluster groupings dependent on the choice of threshold value.

TABLE II
CONFUSION MATRIX

		Predicted	
		Matches Trend	Does Not Match Trend
Actual	Matches Trend	TP	FN
	Does Not Match Trend	FP	TN

The confusion matrix shown in Table II where the labels are defined as follows. The true positives (TP) are the substations that are correctly predicted as following the national trend. The true negatives (TN) are the substations that are correctly predicted as not following the trend. A false positive (FP) is a substation predicted as following the trend when it is not and a false negative (FN) is a substation that is predicted to not follow the trend when it does. Therefore the overall accuracy (ACC) is defined as [6]:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

In addition, the true positive rate (TPR) in Eq. 3 identifies the ratio of substations that match the trend that are correctly identified. Similarly, the false positive rate (FPR) in Eq. 4 identifies the ratio of substations that do not follow the trend that are correctly identified.

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

Lastly, a measure called Matthew's correlation coefficient (MCC) in Eq. 5 is used to account for the biases present in the other metrics by using true positives and negatives and false positives and negatives. MCC ranges from 0 to 1 and a higher value denotes better results. It is analogous to Pearson's product moment correlation coefficient as a measure of fit between the predicted and actual cases in the binary confusion matrix [9].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

In Table III, a few different threshold values are recorded and which roughly correspond to a value taken every 50 substations along with their corresponding ACC, TPR, FPR and MCC. It is evident that in all cases, ACC is a poor measure of how well this method performs, especially in cases where the selected threshold is too high or low. Similarly, TPR and FPR can be misleading because of the nature of the method. In the cases where TP, FP, FN and TN are all nonzero, MCC is the most unbiased metric of the performance of this method.

Ultimately, it is left to the judgment of the network engineers as to how much error is acceptable for the national trend to be used as a proxy for a substation trend. Even the best recorded MCC in Table III of 0.305 is still relatively low but it still highlights the most problematic substations. Therefore, depending on the requirements of error acceptance, this methodology can be used to determine which substations follow the national trend or which substations are the most problematic or a combination of both.

TABLE III
ERROR METRICS

P=30%				
	ACC	TPR	FPR	MCC
T = 0.0035	0.892202	0	0	NaN
T = 0.0046	0.775229	0	0	NaN
T = 0.0054	0.669725	0.053691	0.010453	0.130771
T = 0.0064	0.56422	0.050251	0.004219	0.146207
T = 0.0076	0.522936	0.231076	0.081081	0.198556
T = 0.0093	0.65367	0.654485	0.348148	0.285301
T = 0.0127	0.733945	0.773639	0.425287	0.305018
T = 0.0204	0.915138	1	1	NaN
T = 0.0928	0.997706	1	1	NaN
P=50%				
	ACC	TPR	FPR	MCC
T = 0.0035	0.892202	0	0	NaN
T = 0.0046	0.777523	0.061224	0.014793	0.123593
T = 0.0054	0.674312	0.187919	0.073171	0.172317
T = 0.0064	0.600917	0.336683	0.177215	0.183437
T = 0.0076	0.610092	0.772908	0.610811	0.175523
T = 0.0093	0.690367	1	1	NaN
T = 0.0127	0.800459	1	1	NaN
T = 0.0204	0.915138	1	1	NaN
T = 0.0928	0.997706	1	1	NaN
P=70%				
	ACC	TPR	FPR	MCC
T = 0.0035	0.892202	0	0	NaN
T = 0.0046	0.754587	0.204082	0.085799	0.156329
T = 0.0054	0.538991	0.644295	0.515679	0.122881
T = 0.0064	0.456422	1	1	NaN
T = 0.0076	0.575688	1	1	NaN
T = 0.0093	0.690367	1	1	NaN
T = 0.0127	0.800459	1	1	NaN
T = 0.0204	0.915138	1	1	NaN
T = 0.0928	0.997706	1	1	NaN

VII. CONCLUSION

This paper analyses in-house customer categorisations attributed to substations by way of PCA and clustering. By using PCA, domestic urban, domestic rural and commercial customers can be successfully grouped in the principal component space which follows intuition. By using a biplot, it is possible to achieve a visualisation about the relationship between the PC scores of the substation and the customer variables. Additionally, through the use of clustering and external data, substations with certain PC scores can be identified to be more problematic than others. This provides a basis for looking at the properties of certain customers that deviate the most from the national trend. Of course since this data is unique to the sponsor, it relies on the accurate assessment of customers connected to the distribution network in their respective postal sectors. Therefore this work can be furthered by using more publicly available demographic databases. Furthermore, the shape of load profiles should

also be investigated as it pertains to the distinction between customer groups in this investigation.

REFERENCES

[1] G. Chicco, R. Napoli, F. Pigilione, P. Postolache, M. Scutariu and C. Toader, "Emergent electricity customer classification", *IEE Proceedings - Generation, Transmission and Distribution*, Vol. 152, No. 2, 2005.

[2] Barney Pitt and Daniel S. Kirschen, "Application of Data Mining Techniques to Load Profiling", *Proceedings of the 21st 1999 IEEE International Conference*, pp. 131-136, 1999.

[3] National Grid (2014), "Data Explorer", Available: http://www2.nationalgrid.com/UK/Industry-information/Electricity-transmission-operational-data/Data-explorer/

[4] Vera Figueiredo, Fatima Rodrigues, Zita Vale and Joaquim Borges Gouveia, "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques", *IEEE Transactions on Power Systems*, Vol. 20, No. 2, May 2005.

[5] I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 1986.

[6] Mark A. Hall, Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2011.

[7] K.Y. Yeung and W.L. Ruzzo, "Principal component analysis for clustering gene expression data", *Bioinformatics*, Vol. 17, No. 9, pp.763-774, 2001.

[8] David L. Davies and Donald W. Bouldin, "A clustering separation measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.

[9] DMW Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation", *Journal of Machine Learning Technologies*, Vol. 2, No. 1, 2007.